An Introduction to caGrid Technologies and Data Sharing

Prepared for the caBIG® Community

by the

Documentation and Training Workspace

in cooperation with the

Data Sharing and Intellectual Capital Workspace

and the

caGrid Knowledge Center

May 2010

Table of Contents

1. Introduction

This document[1] presents a high-level overview of a set of technologies called caGrid[2], available from the **National Cancer Institute's Center for Biomedical Informatics and Information Technology (NCI-CBIIT)**. caGrid is a collaborative research and data sharing infrastructure developed under the **cancer Biomedical Informatics Grid (caBIG®)** program. The paper is aimed especially at professionals who may oversee, plan or control data sharing within an institution –for example, an institution's administrators, IRB members, regulatory compliance officers and attorneys, or information technology staff. It also serves as an introduction to caGrid for investigators who may be unfamiliar with the caBIG program but who conduct research under the NIH data sharing mandate and others interested in an overview of the caGrid technologies. Topics covered include the purposes and uses of caGrid technologies, the employment of structured descriptions of data to improve the ways in which it can be exchanged and used, and methods to control access to shared data. Central to caGrid data access control is the establishment of trust relationships between institutions to enable identity verification and access authorization.

The sharing of experimental data been instrumental in advancing biology and medicine, most notably through public databases of gene, protein and genomic sequence data. The caBIG program was conceived to facilitate collaborative research and to support data sharing across the cancer research community. In contrast to many data-sharing efforts, which use centralized databases, caBIG employs a **federated** model, whereby institutions keep their data locally but can share aspects of the data with the public or with chosen collaborators. In particular, sophisticated mechanisms are available to allow sensitive data, such as clinical trial information, to be shared securely with collaborators who present appropriate credentials authenticating their identity. Key to data sharing in caBIG is that institutions retain complete control over the data they share, and also specify how strictly (at what **Level of Assurance**[3]) the identity of the prospective user must be proven.

The focus of caBIG data sharing is translational medicine – the interface bridging medical research and its clinical application. The goal is to expedite the development of treatments relevant to the patient and clinic from discoveries in the laboratory. There is significant support in caBIG, both in the form of software and in the form of organized, collaborative communities, for projects involving clinical trials, tissue banks, pathology, and molecular analysis. Making known the existence and availability of studies, biospecimens, and experiments will facilitate new multi-disciplinary and integrative collaborations.

---

[1] This document is based on "caGrid for Principal Investigators" (2008), which it substantially revises and updates. This new version has an increased focus on the technology and issues surrounding data sharing in caBIG. The document will evolve to address community needs. See Section 12 – Feedback.

[2] This acronym refers to the name "cancer Grid", however the technology is of general utility for any collaborative research and data sharing activity utilizing grid technologies.

[3] See http://csrc.nist.gov/publications/nistpubs/800-63/SP800-63V1_0_2.pdf

Using caGrid technologies, access to data and also to computational analyses are made available to compatible programs via internet-based "**Grid services**" with well-defined interfaces. The **syntax** (how data is represented) and **semantics** (what the data means) of shared data are defined through the use of annotated, centrally registered **Uniform Modeling Language (UML)** models and **controlled vocabularies**. This mechanism is designed to enable a rich level of data reuse and interaction between different programs and data sources across a diverse "virtual" community of users and legacy information systems.

To encourage data sharing, the caBIG program supports the development of supportive open-source applications for **adoption** by participating institutions, e.g. **caTissue Suite** for biospecimen management, **caArray** for storage of microarray gene expression, SNP and other high-throughput genomic data, Clinical Trials Management System (CTMS) for the management of clinical trials and the **National Biomedical Imaging Archive (NBIA)** for the storing of medical images. Using caGrid and other mechanisms, application data can be selectively shared with collaborators or more broadly. The security infrastructure behind this is described in later sections.

For institutions that would like to selectively share data from their existing, well-established data repositories, caBIG provides a second path. Termed "**adaptation**", this path involves extending the existing software so as to provide access to an existing repository using caBIG-compliant methods. This is accomplished by either directly modifying existing code or by developing a new layer which "translates" between the legacy repository and caBIG standards, including caGrid. A hybrid of adoption and adaptation also can be used, whereby data from an existing repository is mapped onto and transferred to a caBIG-supportive application on an ongoing basis.

Extensive support for adoption and adaptation is available through the **caBIG Enterprise Support Network**, which consists of the **caBIG Knowledge Centers** and also of caBIG-licensed **Support Service Providers**. Further information on "adopting" and "adapting" is available in Section 9 below.

2. caBIG Moves Toward Increased Interoperability Using SAIF

At the time of this writing, the mechanisms described in this paper to support interoperability are undergoing substantial revision. The new overall approach is termed the **Services Aware Interoperability Framework (SAIF)** and is based on work undertaken within the **Health Level 7 (HL7)** community. The objective of SAIF is to provide interoperability between computing systems that is not tied to any particular technology. SAIF proposes a structured approach to interoperability, where standard ways of representing data are used from the beginning of the specification process for services development. The caBIG "compatibility" methods previously relied on procedures whereby many shareable data elements were designed by the community on an as-needed basis. A caBIG interoperability implementation guide is under

development for 2010 release. It is anticipated that working towards these goals will involve numerous changes to the use of caBIG technology described in this paper, including movement towards a **services-oriented architecture (SOA)**.

3. caGrid is the Infrastructure for Data Sharing in caBIG

caGrid is the infrastructure that enables secure, federated sharing of data and data analysis capabilities. caGrid extends the open-source **Globus**[4] framework with support for querying remote data sources, vetting identities and implementing data sharing policies, managing data security, and for locating grid-based services. The term caGrid is used to refer both to the software technologies developed with support from caBIG, as well as to the particular production grid deployed for caBIG.

The grid approach addresses key problems encountered in enabling data sharing. As defined in Foster *et al.* (2001), grid computing is "controlled and coordinated resource sharing and problem solving in dynamic, scalable virtual organizations. "Controlled" implies that resource owners retain both authority and responsibility to determine whether other institutions and users can access their data and computational resources. The purpose of the grid is to solve problems that require dynamic interaction among collaborators.

Configuring a computer, server, database, or facility to be a member of a Grid is accomplished by installing and running appropriate software and implementing that Grid's policies and security protocols. The only special characteristics of the particular computers that make up a given grid are that

1. they can find and communicate with each other in a reliable and secure fashion
2. they trust each other.

Point 1 is addressed by requiring grid-based services to **advertise** themselves to a yellow-pages-like directory (in caGrid this is called an "**Index Service**", which is itself a Grid service) so that they can be searched for and found, and by using standard protocols such as XML messages for communication. As for point 2, trust is established through policy and procedural mechanisms formally accepted in signed agreements between the providers of the core Grid infrastructure itself (e.g. security and index services), the owners of Grid services, and by service users. Depending on the requirements of the resource owners, access can be as restrictive or as open as desired.

caGrid data transport uses standard internet protocols. caGrid adds an additional layer to the internet, for example by specifying a particular structure to data communications using XML schemas, and through its provision of authentication and authorization services. Putting a service "on the Grid" means making it conform to the caGrid conventions and specifications and advertising its availability; it does *not* mean that the

[4] http://www.globus.org/

service is hosted by NCI or that any particular data is necessarily "deposited" into any central data base.

4. caGrid Services

caGrid software is used to construct two basic types of services: **data services** and **analytical services**. Unlike a web site, these internet-based "services" are intended to communicate with other applications rather than directly with humans. The distinction between the two types is that a data service provides access to structured data, whereas an analytical service performs specialized calculations. The caGrid infrastructure itself is assembled from Grid services (Figure 1), several of which are described in subsequent sections.

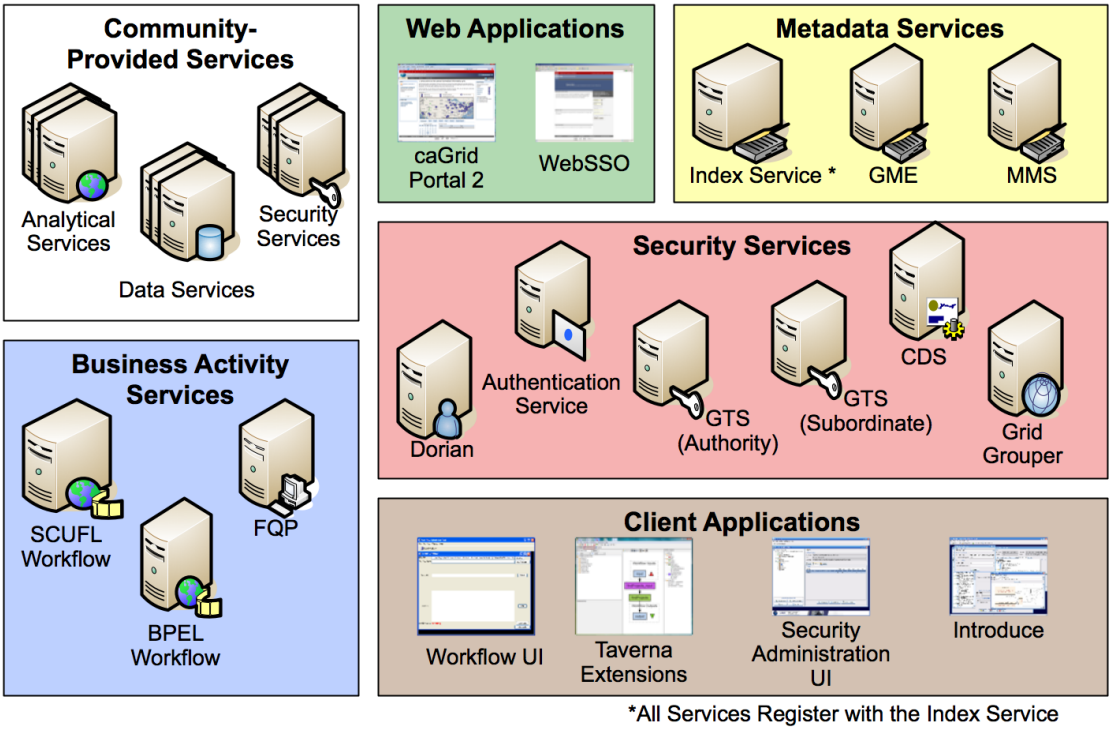

Figure 1- caGrid services

There is generally no requirement to install "Grid" software on the computers of individual users. Most caGrid-based services instead provide a web interface or dedicated client program by which they can be accessed. It is important to again note that data is typically maintained locally.

Data services act as a conduit to databases and other data management systems, such as those acting as repositories for tissue bank and clinical trials records, or for the results of microarray gene expression experiments. Queries are typically written in the "caGrid

Query Language" (**CQL** [5]). CQL was developed to allow creation of queries using "**object-oriented**" languages. (In an object-oriented system, data is bundled together with software methods to manipulate it in what is called an object).

Analytical services perform calculations on a provided data set and return a result. Examples include hierarchical clustering, gene copy number variation analysis (GISTIC), and gene expression profile correlation (GeneNeighbor). The format in which they expect to receive and return data, and any parameters required for controlling the calculation, are all specified in interface documentation[6]. As with all Grids, the intent is that data should be housed and calculations performed in the places most appropriate for each. Data may be housed at the institution where it was created, while an analytical service might be offered by an institution with expertise in implementing a particular type of algorithm or calculation. Different data and analytical services can then be joined to carry out a particular calculation, with data transferred as needed. The **caGrid Portal** website allows users to directly explore the types and location of caBIG Grid services available.

Invoking a data or analytical service is typically done with "**client**" software[7] provided by the service developer. This client may run in a web browser, as a desktop application, or be a software component that can be added to another program. The client software typically shields the end-user from any of the details of how its queries are performed.

In addition to access to a data service via client software, caGrid data services also allow for direct, **dynamic invocation** of queries, which are built in CQL according to the service's information model. caGrid provides a generic CQL software client for managing dynamic queries. Such queries will become useful in constructing **workflows** "on-the-fly"- that is, discovering data sources and constructing a series of steps that make use of the data, all without downloading any specific client software for the particular data services being used.

The distinction between data services and analytical services is to some extent more one of definition than technology, as both can implement the same kinds of pre-defined operations[8]. However, analytical services operate on data only transiently, while data services maintain it long-term and support CQL-based queries.

It is important to again note that the caGrid architecture is moving toward a Services Oriented Architecture (SOA) so that services can be more highly interoperable and reusable. The migration to SOA will be occurring in conjunction with the adoption of the SAIF (see Section 2) and its underlying Enterprise Conformance and Compliance Framework (ECCF) and other interrelated SAIF frameworks to specify and trace requirements through the development of services.

---

[5] Also referred to as "Common Query Language".
[6] This published, public interface is referred to as an "**application programming interface**", or **API.**
[7] This client is typically distributed as a Java "jar" file.
[8] caArray from version 2.3.0 onwards provides such a predefined grid data query interface.

5. Transport of Data

The caGrid-based infrastructure currently offers several mechanisms for transporting data. The first is suitable for regular text-based records and uses **XML (Extensible Markup Language)** messages. XML is a markup language similar to the Hypertext markup Language (HTML) used for web page display, but devoted to data interchange. XML is text-based and is independent of any particular type of computer or operating system (platform-independent). A major feature of XML is that both computers and (some) humans can interpret it. User-defined tags describe each hierarchical level and item of the data.

For transferring large amounts of numeric or binary data, *e.g.* raw gene expression or image file data, other mechanisms have been developed[9]. One such method is called the caGrid **Transfer Service**, which has been successfully used by several projects. This method transfer files in their original format. A second method, the **Enumeration Service,** allows large data sets to be retrieved in sequential chunks of limited size.

6. Interoperability

As already mentioned, making applications and data sources more useful by facilitating their ability to interoperate has from the start been a key goal of the caBIG program. An important part of the strategy for realizing interoperability has been to encourage the reuse of standardized data objects by different projects. For example, if an application or an analytical service accepts a specific object type provided by a given data service, it is possible to directly link them.

As reuse of standard data objects increases, software tools for constructing workflows will be able to dynamically discover and join together various compatible caGrid services to carry out a particular task, without *a priori* knowledge of the format and meaning of the data on the part of the person designing the overall task. The caBIG program's transition to specifying interoperability using the SAIF framework will enable computable interoperability as semantics are more explicitly specified and instantiated in services.

At the practical level, two types of interoperability can be described. These are "**syntactic**" and "**semantic**" **interoperability (**Figure 2**)**.

[9] XML representation of numeric data can lead to a large increase in the size of the data being transferred, and for large data sets places corresponding demands on the resources of the receiving machine.

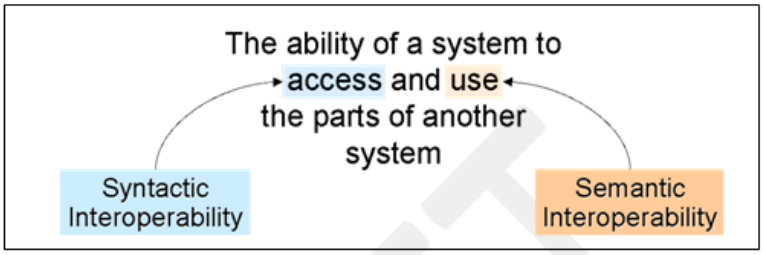

Figure 2 – Syntactic and Semantic Interoperability

- **Syntactic interoperability** means that any two interacting software components agree on the number, type and order of pieces of information to be exchanged. Simply put, the program making a call and the one receiving it fit together properly, with all needed information present and in the correct format and order.

- **Semantic interoperability** means that the programs on the two sides of an exchange have the same understanding of what the items being exchanged represent. Semantic interoperability is achieved by creating unambiguous descriptions of the information exchanged through the use of controlled vocabularies and definitions of **concepts** being used.

The annotations that describe the syntax and semantics are termed **metadata**, that is, data about data.

For further information see the caBIG Core Concepts page at https://cabig.nci.nih.gov/overview/caBIG_core_concepts.

7. Key Compatibility Guidelines

During its initial phase, caBIG developed guidelines designed to facilitate interoperability of applications and services. Although these compatibility guidelines will be superseded by the new SAIF implementation (Section 2), the principles are general and may continue to apply to existing and some new applications for some time. The guidelines address four specific topics and are summarized below.

The first three concern "semantic" interoperability:

- ***Information/Data Models*** – these describe the data elements and the relationships between them in a system. In caBIG, these models are created using UML.
- ***Controlled Vocabularies*** –Descriptive terms used in the UML models are drawn from the caBIG Enterprise Vocabulary Services (**EVS**) system, which is curated under NCI-CBIIT supervision. The terms used in the EVS come from a variety of sources including several licensed terminologies made available, by permission, to the broader NCI and biomedical community, while providing notice of and enforcing some limitations on their use.
- ***Common Data Elements*** – these standardize the types and permissible values of data, and are the basic element of exchange between different systems.

The fourth topic concerns "syntactic" interoperability:

- ***Programming and Message Interfaces*** – the technical details involved in exchanging commands and data between systems. The syntax of data described in models and transferred over the Grid is defined by XML schemas[10], and the data is transported wherever practical as XML text.

The detailed descriptive information about data objects (both the syntactic and the semantic definitions) is stored into the **Cancer Data Standards Repository (caDSR)**. The caDSR is a highly structured database, curated under the supervision of NCI-CBIIT. When creating new applications or services under this framework, software designers should consult the caDSR for existing, reusable CDEs, in order to facilitate the future ability to interoperate with other software and prevent rework when data exchange is envisioned.

8. The caGrid Security Infrastructure

A need for data security can arise from a number of sources. Cancer Centers may house clinical or experimental data that they wish to share only with a limited, authorized list of collaborating investigators. Sensitive data may need to be protected while in transit from a database to the authorized investigator using it, or while en route between an investigator and an analytical service. Even de-identified clinical data may require protection if it is regarded as proprietary or access is restricted for any reason by its steward. A Center or other user of caGrid may be subject to various federal, state or local regulations or other requirements such as the **Health Insurance Portability and Accountability Act (HIPAA)**. The Data Sharing and Intellectual Capital Workspace (DSIC) and Knowledge Center have developed a Data Sharing and Security Framework and associated decision support tools to facilitate assessing the sensitivity of data. For further information on the caGrid Enterprise Security policies see the policy documents on the Enterprise Security Program wiki.

When one attempts to use a caGrid service that runs securely and requires login, two separate levels of verification may be instituted. In the first, the identity of the user is established – that is, the identity is **authenticated**. In the second, the service (maintained by the data owner) may check whether that particular authenticated user is **authorized** to gain access to a requested data set or portion of a data set.
A collection of caGrid services called **GAARDS** (Grid Authentication and Authorization with Reliably Distributed Services) provides the core security capabilities needed to securely exchange data between Grid services and applications. Here we introduce some of its basic components and describe the general purpose of each. Some of the following material may be more informative to those in more technically oriented areas such as an Informatics Department, IT Security Group or Chief Information Office. Additional support in understanding caGrid security services may be obtained by working with one or more of these local offices and also by working with the caGrid Knowledge Center.

[10] The operational copies of the XML schemas used by grid services are stored in the GME.

Authentication

Authentication refers to proving the identity of a user – that the user is who he or she claims to be. Although many institutions may use a simple username/password combination for user authentication, more elaborate methods are available for protecting access to secure data, such as those using biometrics or electronic keys.

The caBIG program has adopted the **National Institute of Standards and Technology's** **Electronic Authentication Guideline** intended for federal agencies implementing electronic authentication. The Electronic Authentication Guideline defines four increasingly stringent **Levels of Authentication Assurance** (LOA) from Levels 1 to 4, in terms of the consequences of authentication errors,the types of acceptable credentials for each Level and the identity vetting process for each Level of Assurance. With each successive level of assurance, the confidence that can be placed in the authentication increases. This guideline defines the technical requirements for each of four levels in the areas of identity proofing, registration, tokens, authentication protocols and related assertions. The GAARDS component **Dorian** provides support for the first three of the four increasingly stringent "Levels of Assurance" of authentication.

NCI-CBIIT maintains a national-level installation of Dorian[11] to provide authentication services to the caBIG Production Grid (caGrid). It can support two types of authentication, either (1) direct or (2) delegated.

(1) Under direct authentication, the user sets up a Dorian-based caGrid account through NCI-CBIIT after establishing his or her identity and affiliation. These accounts are maintained in the national-level installation of Dorian. During caGrid login, the user then provides a username and password directly to Dorian. In this scenario, Dorian directly verifies the user's information (that the username and password are accurate) and issues a **Grid credential**[12] to the user.

(2) Under delegated authentication, a user's institutional account is used to login to the Grid. In this case, authentication is performed locally at the user's institution, and an **identity token** is sent on to Dorian, which validates it and subsequently issues a Grid credential to the user. This requires that the IT staff at the local institution or department install and configure a specific software package (the GAARDS **Authentication Service**) that can access the local authentication system. In a typical example, a department's local authentication system may be a central **Lightweight Directory Access Protocol** (**LDAP)** service that maintains all usernames and passwords. Following local authentication through the Authentication Service, a secure, trusted identity token is forwarded to Dorian. Under this procedure there must be a **trust agreement** in place between the local institution and caGrid as represented by Dorian. That is, Dorian must have been configured to accept a credential from the institution's authentication service. This second method allows an institution to directly provide relevant personnel with access to caBIG Grid services.

[11] Institutions are free to install local copies of Dorian for internal or collaborative use.
[12] The credential is known as an X.509 certificate, created using public key encryption methods (PKI).

The first route, authenticating using the national-level Dorian service, might be appropriate for users at smaller institutions that do not have the staff or resources to implement local caGrid authentication. The second, building on local authentication, might be more attractive to larger centers as caGrid use becomes more common. This would allow users to leverage their existing credentials to authenticate to the Grid, allowing easy transition between every-day applications such as desktop login, E-mail and Grid applications.

Once the identity of a user has been authenticated, Dorian issues a Grid credential to the user, which can be presented (via software) to other caGrid services as proof of identity. A Grid credential is required to use any secure service available on the Grid. This credential consists of a X.509 certificate and private cryptographic key that can be used to convey the identity of the person holding it, and to encrypt communications with remote Grid services. In caGrid, setting up a secure communication channel with a service involves mutual authentication: both parties verify each other's identity[13].

In a final step the actual caGrid service the user wishes to access validates the user's Grid credential. To do this it submits the credential to a **Grid Trust Service** **(GTS)** (another component of the GAARDS package). The Grid Trust Service maintains a list of all trusted Certificate Authorities, and verifies that the presented Grid credential is real and signed by one of those trusted sources (*e.g.* Dorian).

Authentication Levels - Policy and Deployment

A **Certificate Authority (CA)** issues trusted digital certificates. Examples of publically available CAs are those provided by firms such as VeriSign, Entrust and Verizon and those provided by organizations such as the InCommon Federation and the SAFE[14] Biopharma Association. Dorian acts as a CA to enable existing identity providers to be integrated into the Grid (it is also possible to use traditional commercial CAs). Once an identity provider has been registered with the Dorian CA, its users can access the Grid with their existing credentials. Figure 3 below illustrates how GAARDS can be deployed in a production environment in compliance with the Electronic Authentication Guideline. Authentication Services complying with a given Level of Assurance are registered with Dorians complying with that same level. In turn, Dorians are registered with the Grid Trust Service at the Electronic Authentication Guideline Level of Assurance with which they comply.

[13] https://cabig-kc.nci.nih.gov/CaGrid/KC/index.php/Create_a_Secure_Data_Service_using_CSM_for_Data-Level_Authorization
[14] "Signatures and Authentication for Everyone"

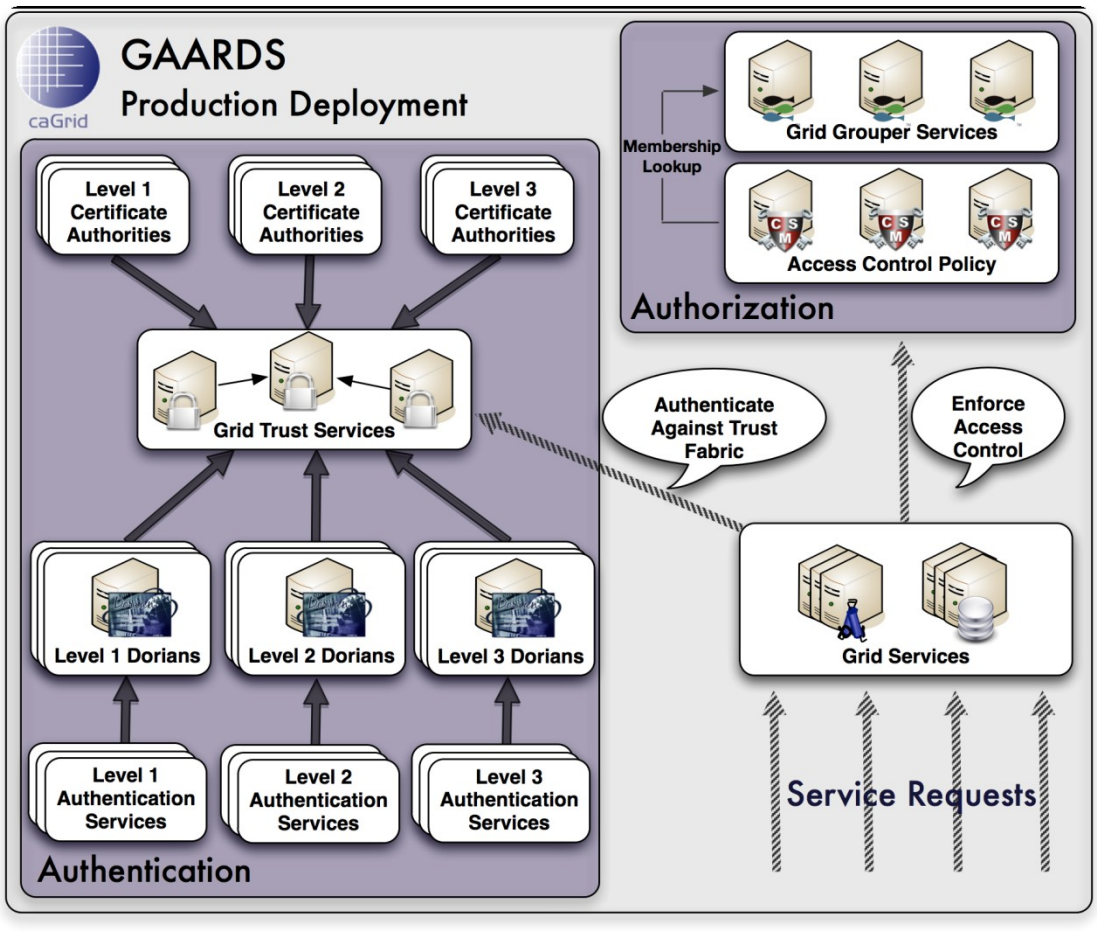

Figure 3 - Authentication and authorization in GAARDS

When clients make calls (requests) to secure services, both sides must authenticate themselves using their X.509 credentials. The credentials are presented to the Grid Trust Service to verify that they were issued and signed by a CA at the level of assurance required by that service. For example, a service may specify that it will accept Level 2 and Level 3 credentials. Since all trusted CAs are registered with the GTS, the service needs to verify that a certificate presented is signed by a level 2 or level 3 CA that is registered to the GTS. If the certificate presented is signed by a CA that is not in the GTS or is signed by a CA not registered at level 2 or level 3 (i.e. may be unregistered or registered at level 1), the service request is rejected.

Authorization

Certain caGrid data services (and data stewards) may need to restrict access to only certain individuals, or restrict the types of operations that different groups of users can perform or restrict the data they can access. For example, different levels of access might be granted to the principal investigator, the statistician, and the technician involved in a particular study.

The most important aspect of authorization is that it is performed locally, by the service itself. Thus the institution hosting a service always controls access to its data and is responsible for any regulated controls over the access to that data such as may be applicable from HIPAA and Institutional Review Board or other state or local privacy requirements.

The increasingly detailed levels of authorization can be termed "service-level", "operation-level" and "fine-grained" or "data-level".

All three levels of authorization can be provided by the GAARDS component **Grid Grouper**. Grid Grouper provides a means to define and manage group membership. The administrators of a given caGrid service can define a group and control what access privileges attach to membership in that group. They also control which caGrid users are added to the group. A Grid service can check with Grid Grouper "behind the scenes" to see if the action is allowed for that user. It is entirely up to the particular service what privileges it grants to a particular user based on the group membership.

Successful authorization at the "service-level" grants access to the Grid service as a whole. At the second, or "operation-level" of authorization, access can be controlled to particular operations available on a Grid service. This level of access can be implemented by setting up separate groups of users (*e.g.* again using Grid Grouper) with different access privileges. The Grid service is then responsible for checking the group to which a user belongs and granting or denying access to particular operations based on this.

A more "fine-grained", role-based access control policy can be integrated locally into a data service maintained and enforced through use of the **Common Security Module (CSM) that provides authentication and authorization services**. Access control policy can be specified based on user identities, locally defined groups, or groups defined in Grid Grouper. The CSM can be used to control access to portions of a data service's information model, or specific sets of data stored in the data service. Access to specific operations or data subsets can be specified and granted as "**roles**". For example, one role might consist of the permissions granted to the principal investigator. The CSM will also soon be available as an invokable service so that it need not be integrated or embedded directly into a data service's software code but may be called upon by any data service as needed.

An access control framework is currently under development to help institutions or caBIG participants define their authorization model. The framework will provide a structured methodology (termed a "standard role engineering process") for ensuring that critical principles (e.g., separation of duties, use of least privileges necessary) in defining roles and permissions are met. It will also define common reference roles for use within particular contexts (e.g., study coordinator in clinical trials) so that applications in the same or similar work domains can use the same roles to designate similar responsibilities. By applying a standard role engineering process and using reference common roles in the

same domain, a high level of assurance of automated control and security can be achieved while maintaining local authorization of caGrid services.

For more information on policies concerning data sharing and security in caBIG please visit the DSIC Knowledge Center and the caGrid Knowledge Center and the caGrid Enterprise Security Program wiki.

9. Making A Data or Analytic Service Available via caGrid

We now turn to practical aspects of how caGrid technologies can be brought into use by individual groups or institutions to share their data freely or within a secure collaborative framework. In some cases, a central data repository may be available, such as the NCI-maintained installation of caArray. However, the caBIG project is aimed more at creating and linking distributed, federated resources.

There are two common approaches within caBIG for sharing locally maintained data and resources. The first is to use ready-made caBIG-compatible software. This is termed adoption of caBIG software. The second approach is to create a caGrid/caBIG compatible interface to an existing data resource or application. This is called adaptation – that is, you adapt what you already have to the requirements of caBIG. Of course, new resources can be created which are designed from the beginning to be caBIG compatible. Adoption and adaptation are further explained below.

Adoption - Install a copy of already available, caGrid-enabled software

If a caBIG-compatible software package exists which meets your requirements, adopting it may be very advantageous. Scenarios include

1. You have no existing system, and the caBIG application meets the stated needs
2. There is an existing sytem, and its data can be loaded, once or periodically, into the caBIG compatible system.

You could expect to need at least the following support:

- System administrator with web application server experience, including experience in IT security, basic understanding of digital certificate concepts and group-based authorization concepts,
- Possibly the help of a database administrator to set up and port data to the new repository, and
- A technician, student or other personnel comfortable with computer applications to actually use the new software in a research setting.

The time required to adopt a particular application will include installation, loading of data and training, and will vary depending on a number of variables such as the complexity of the application being deployed, the technical experience of the staff and

the organizational dynamics and control of the infrastructure. The most important variable may be the extent to which legacy data needs to be mapped and transformed for use with the new system.

The Enterprise Support Network can provide resources and guidance to help complete your installation. For further information on "Getting on the Grid" please also see the site https://cabig.nci.nih.gov/working_groups/DSIC_SLWG/sharable/cagrid_overview.

Adaptation - Build caGrid interfaces to your existing database or software.

You may already have a data repository or application that you wish to continue using but that you would also like to make available to others through using caGrid. Or you may be starting on a new project and wish to utilize standards to ensure caBIG-compatibility from the start. Either path will involve effort in object modeling, software design and in programming. In addition to the requirements listed for installing a service (see Adoption, above), you may need a developer or developer team experienced with the following:

- Current web application server technologies,
- Modeling a service, *e.g.* using UML, and
- Filling out code in the service framework created through using the **Introduce** toolkit. The Introduce toolkit (which is developed as part of the caGrid suite of programs) greatly simplifies the work that needs to be done in creating the basic framework.

As with any software development task, time must be allocated for design, development, testing and training, and will depend greatly on the previous experience of staff with the technologies. Details of creating information models as required for transmitting data using caGrid are provided in the next section.

The caBIG Enterprise Support Network can be consulted to help you complete this process. Also see Lesson 6 in the caBIG Essentials Overview training module, "Adapting a Tool to be caBIG Compatible: Overview".

Creating and Reusing Information Models

To promote interoperability and reusability of caBIG data, most information (in the form of the data objects) exchanged on the caBIG Grid should be described and annotated in detail in the form of **information models**. Central to this effort is that such data models be reused as much as possible by different applications and resources. This is not an easy requirement and much effort is being given to making it easier to find and use existing data models, and to create an overall framework which will guide the creation of new, reusable data models (*e.g.* the Life Sciences Data Analysis Model, or **LS-DAM,** and more recently the adoption of SAIF (Section 2).

Information models are typically developed by the programming team for a particular project, utilizing specialized software[15]. As already noted above, you may be developing an information model to adapt an existing data resource to caBIG, or you may be creating a new resource from scratch. The best approach is to examine the project requirements and find or create information models before code development begins. Existing models can be found in the caDSR.

The models are developed using Unified Modeling Language (UML). This is a task for software professionals, and requires knowledge of object-oriented software design principles.

For further information see the caCORE Training site.

10. Workflows in caGrid

A workflow is a sequence of steps carried out to accomplish a set task. It might involve steps that are complex or lengthy, and would be difficult and unproductive for a person to repeatedly execute. Workflows can also be used to allow non-experts to carry out sophisticated tasks, by pre-packaging the steps and settings, thus reducing learning time and errors.[16]

A popular system for composing workflows is called **Taverna**, which provides a graphical interface for workflow creation and execution. It was developed as part of the **MyGrid** project in the United Kingdom, where it has found support among the Bioinformatics community, among others.

As a result of cooperation between caBIG and MyGrid, caGrid-based services can be integrated into workflows composed in Taverna, which now supports the caGrid security architecture. Support for running Taverna workflows is also being added to the caGrid Portal (as the "**Taverna Workflow Portlet**"). In addition, a Taverna plugin (**CQL Builder**) has been created that allows simple creation of CQL queries.

As touched upon in Section 4, a goal of caBIG is to discover, understand, and match up caGrid services in a relatively automatic fashion to carry out the desired steps of a workflow. The syntactic and semantic information needed for this is stored in the models and their metadata, which reside in repositories such as the caDSR. Additional support from local IT or Informatics staffs or CIO offices in addition to consulting with the caGrid Knowledge Center may be useful to understanding the caGrid services.

[15] Examples are Enterprise Architect and ArgoUML.
[16] Yolanda Gil, "Semantic Workflows for caBIG, Metadata Meets Computational Workflows", presentation at the October 2009 Architecture/VCDE Face-to-Face meeting in Atlanta.

11. Policy and Technology in Future Extensions of the caBIG/caGrid Implementation

The caGrid infrastructure continues to evolve to respond to developments in technology and user needs. Authentication in a federated environment is complex. While caGrid implements NIST standards for vetting the identity of credential holders, challenges remain for institutions to accept the credentials of unknown individuals for purposes of sharing data. The implementation of role- and attribute-based authorization is also complex. Efforts to define a more fine-grained access control policy and infrastructure, essential to scaling data sharing across a large set of users, are underway. Finer-grained access to, for example, individual aspects of studies may be needed. As caGrid evolves toward a Services Oriented Architecture, and increasingly adopts international interoperability standards, more change will result. The ever-changing landscape of technology creates both opportunities and challenges. Keeping up requires balancing competing needs and creating effective tools to fill those needs. Standards are evolving and becoming increasingly sophisticated in response.

Discussion continues in these areas as community feedback is received.

12. Feedback

Further questions on the topics covered in this document should be submitted to the appropriate subject forums at the caGrid Knowledge Center.

Feedback on this document, as to its usefulness or suggestions for improvements, can also be submitted to the Suggestions Forum at that site.

13. Acknowledgements

First, I would like to thank all those who worked on this paper for the time, effort and knowledge they contributed through its numerous revisions – this paper is truly a community effort. I would also like to thank Jenny Tucker, leader of the caBIG Documentation and Training Workspace, for strongly supporting the creation of both this paper and its predecessor, "caGrid for Primary Investigators", and Justin Permar of the caGrid Knowledge Center for his advice and technical expertise on both papers. I am very grateful to Marsha Young of the caBIG Data Sharing and Intellectual Capital Workspace for organizing and guiding this project, for making substantial contributions to the content, and for serving as its editor. Finally I wish to thank the members of the caBIG community who reviewed and provided invaluable comments on this paper - and to those who will in the future provide suggestions to improve it. – KCS.

14. Acronyms

CA	Certificate Authority

caBIG®	Cancer Biomedical Informatics Grid
caDSR	Cancer DataStandards Repository
CQL	caGrid Query Language or Common Query Language
CSM	Common Security Model
DSIC	caBIG Data Sharing and Intellectual Capital Workspace
EVS	Enterprise Vocabulary Services
GAARDS	Grid Authentication and Authorization with Reliably Distributed Services
GISTIC	Genomic Identification of Significant Targets in Cancer
GTS	Grid Trust Service
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level 7
LDAP	Lightweight Directory Access Protocol
LS-DAM	Life Sciences Domain Analysis Model
NBIA	National Biomedical Imaging Archive
NCI	National Cancer Institute
NCI-CBIIT	National Cancer Institute – Center for Biomedical Informatics and Information Technology
SAFE (BioPharma Association)	Signatures and Authentication for Everyone
SAIF	Services Aware Interoperability Framework
SOA	Services Oriented Architecture
UML	Unified Modeling Language
XML	Extensible Markup Language

15. References and links to further reading

Foster I, Kesselman C., Tuecke S. (2001). The Anatomy of the Grid: Enabling Scalable Virtual Organizations (International J. Supercomputer Applications, **15**(3)

Langella S, Hastings S, Oster S, Payne P, Siebenlist F (manuscript in preparation) Authentication and Authorization in Cancer Research Systems.

OMB Memorandum 04-04, E-Authentication Guidance for Federal Agencies: http://www.whitehouse.gov/OMB/memoranda/fy04/m04-04.pdf

NIST Electronic Authentication Guideline (2006)
http://csrc.nist.gov/publications/nistpubs/800-63/SP800-63V1_0_2.pdf

The Globus Grid project: http://www.globus.org/

Enterprise Conformance and Compliance Framework" (ECCF). http://ec2-174-129-196-76.compute-1.amazonaws.com/mediawiki/index.php/Main_Page

Yolanda Gil, "Semantic Workflows for caBIG, Metadata Meets Computational Workflows", presentation at the October 2009 Architecture/VCDE Face-to-Face meeting in Atlanta.

Additional information on caBIG, caGrid, and the topics introduced in this document can be found at the locations shown in the table below. In particular, much of the material covered in this document is treated in more detail in the caBIG Essentials (https://cabig.nci.nih.gov/concepts/essentials) presentation.

Name	URL
caArray	https://cabig.nci.nih.gov/inventory/data-resources/caarray/
caBIG caCORE Training (incl. caDSR *etc*)	http://ncicb.nci.nih.gov/NCICB/training/cadsr_training
caBIG Community website	https://cabig.nci.nih.gov
caBIG Community website – "Getting Connected"	https://cabig.nci.nih.gov/getting_connected
caBIG Compatibility	https://cabig.nci.nih.gov/sharable/compatible
caBIG Compatibility and Certification	https://cabig.nci.nih.gov/guidelines_documentation
caBIG Compatibility Guidelines	https://cabig.nci.nih.gov/guidelines_documentation/compat_v3/
caBIG Core Concepts	https://cabig.nci.nih.gov/overview/caBIG_core_concepts
caBIG Data Sharing	https://cabig.nci.nih.gov/working_groups/DSIC_SLWG/data_sharing_policy
caBIG Documentation and Training Workspace	https://cabig.nci.nih.gov/working_groups/Training_SLWG/
caBIG Essentials presentation (Flash and PowerPoint)	https://cabig.nci.nih.gov/concepts/essentials
caBIG Enterprise Support Network	https://cabig.nci.nih.gov/esn
caBIG Enterprise Vocabulary Services	https://cabig.nci.nih.gov/concepts/EVS
caBIG Knowledge Centers	https://cabig.nci.nih.gov/esn/knowledge_centers
caBIG Support Service Providers	https://cabig.nci.nih.gov/esn/service_providers
caGrid and Infrastructure Overview	https://cabig.nci.nih.gov/sharable/sharable/cagrid_overview

caGrid Enterprise Security Program	caGrid Enterprise Security Program wiki.
caGrid Knowledge Center	https://cabig-kc.nci.nih.gov/CaGrid/KC
caGrid Knowledge Center Forums	https://cabig-kc.nci.nih.gov/CaGrid/forums/
caGrid	https://cabig.nci.nih.gov/workspaces/Architecture/caGrid
caGrid development information	http://www.cagrid.org
caCORE Common Security Module (CSM) – Wiki	https://wiki.nci.nih.gov/display/caCORE/Common+Security+Module+(CSM)
caCORE CSM – Tool page	https://cabig.nci.nih.gov/tools/CSM
caCORE CSM – Data-level authorization for caGrid services	https://cabig-kc.nci.nih.gov/CaGrid/KC/index.php/Create_a_Secure_Data_Service_using_CSM_for_Data-Level_Authorization
caDSR	https://cabig.nci.nih.gov/concepts/caDSR
Credential Delegation Service	https://cabig-kc.nci.nih.gov/CaGrid/KC/index.php/CredentialDelegationService
Dorian	https://cabig-kc.nci.nih.gov/CaGrid/KC/index.php/Dorian
ECCF	https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/SI_Conop_Role_Of_ECCF
GAARDS security suite information	https://cabig-kc.nci.nih.gov/CaGrid/KC/index.php/GAARDS
Globus Project	http://www.globus.org/
Grid Grouper	https://cabig-kc.nci.nih.gov/CaGrid/KC/index.php/Grid_Grouper
Grid Trust Service	https://cabig-kc.nci.nih.gov/CaGrid/KC/index.php/GridTrustService
Introduce	https://cabig-kc.nci.nih.gov/CaGrid/KC/index.php/Introduce
MyGrid	http://www.mygrid.org.uk/
NCI-CBIIT	http://ncicb.nci.nih.gov/
SAFE Biopharama Association	http://www.safe-biopharma.org/
Taverna	https://cabig.nci.nih.gov/tools/taverna

16. Credits

Principal Author

Kenneth C. Smith, Ph.D.
(caBIG Documentation and Training Workspace)
Joint Centers for Systems Biology
Columbia University
New York, New York

Contributors/Editors:

Brian Davis, PhD.
(caBIG Vocabulary and Common Data Elements Workspace)
3rd Millennium, Inc.
Waltham, Massachusetts

Nina Kudszus, JD.
(caBIG Data Sharing and Intellectual Capital Workspace)
Booz Allen Hamilton
Rockville, Maryland

Stephen Langella, MS
Inventrio
Columbus, Ohio

Ken Lin, MS, CISSP
Booz Allen Hamilton
Rockville, Maryland

Justin Permar, MS
(caBIG caGrid Knowledge Center)
Software Research Institute
Center for IT Innovations in Healthcare
The Ohio State University
Columbus, Ohio

Jennifer Tucker, PhD.
(caBIG Documentation and Training Workspace)
Otto Kroeger Associates
Fairfax, Virginia

Marsha Young, JD
(caBIG Data Sharing and Intellectual Capital Workspace)
Booz Allen Hamilton
Rockville, Maryland